Why and How Does Consciousness Seem the Way it Seems?

Daniel C. Dennett

Are-expression of some of the troublesome features of my oft-caricatured theory of consciousness, with new emphases, brings out the strengths of the view and shows how it comports with and anticipates the recent introduction of Bayesian approaches to cognitive science.

Keywords
Bayes | Consciousness | Hume | Inversion | Qualia | Transduction

Author

Daniel C. Dennett
daniel.dennett@tufts.edu
Tufts University
Medford, MA, U.S.A.

Commentator

David Baßler
davidhbassler@gmail.com
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

1 Introduction

People are often baffled by my theory of consciousness, which seems to them to be summed up neatly in the paradoxical claim that consciousness is an illusion. How could that be? Whose illusion? And would it not be a *conscious* illusion? What a hopeless view! In a better world, the principle of charity would set in and they would realise that I probably had something rather less daft in mind, but life is short, and we'll have one less difficult and counterintuitive theory to worry about if we just dismiss Dennett's as the swiftly self-refuting claim that consciousness is an illusion. Other theorists, including, notably, Nicholas Humphrey (2006, 2011), Thomas Metzinger (2003, 2009) and Jesse Prinz (2012), know better, and offer theories that share important features with mine. I toyed with the idea of trying to re-offer my theory in terms that would signal the areas of agreement and disagreement with these welcome allies, but again, life is short, and I have found that task simply too much hard work. So with apologies, I'm going to restate my position with a few new—or at least newly emphasized—wrinkles, and let them tell us where we agree and disagree.

I take one of the usefully wrong landmarks in current thinking about consciousness to be Ned Block's attempt to distinguish "phenomenal consciousness" from "access consciousness."

His view has several problems that I have pointed out before (Dennett 1994, 1995, 2005; Cohen & Dennett 2011), but my criticisms have not been sufficiently persuasive, so I am going to attempt, yet again, to show why we should abandon this distinction as scientifically insupportable and deeply misleading. My attempt should at least help put my alternative view in a better light, where it can be assayed against the views of Block and others. Here is the outline, couched in terms that will have to be clarified and adjusted as we go along:

1. There is no double transduction in the brain. (section 1)
   Therefore there is no second medium, the medium of consciousness or, as I like to call this imaginary phenomenon, the *ME*dium.
   Therefore, qualia, conceived of as states of this imaginary medium, do not exist.
2. But it seems to us that they do. (section 2)
   It seems that qualia are the source or cause of our judgments about phenomenal properties ("access consciousness"), but this is backwards. If they existed, they would have to be the *effects* of those judgments.
3. The seeming alluded to in proposition 2 is to be explained in terms of Bayesian expectations. (section 3)
4. Why do qualia seem simple and ineffable? This is an effect, a byproduct, an artifact of "access consciousness." (section 4)
5. *Whose* access? Not a witness in the Cartesian Theater (because there is no such functional place). (section 5)
   The access of other people! Our "first-person" subjectivity is shaped by the pressure of "second-persons"—interlocutors—to have practical access to what is going on in our minds.
6. A thought experiment shows how even color qualia can be understood as Bayesian projections.

2 There is no double transduction in the brain

The arrival of photons on the retina is transduced thanks to rhodopsin in the rods and cones, to yield spike trains in the optic nerve (I'm simplifying, of course). The arrival of pressure waves at the hair cells in the ear are similarly transduced into spike trains in the auditory nerve, heat and pressure are transduced into yet more spike trains by subcutaneous receptors, and the presence of complex molecules in the air we breathe into our noses is transduced by a host of different transducer molecules in the nasal epithelium. The common medium of spike trains in neuronal axons is well understood, but used to be regarded as a baffling puzzle: how could spike trains that were so alike in their physical properties and patterning underlie such "phenomenally" different phenomena as sight, hearing, touch, and smell? (see Dennett 1978, for an exposure of the puzzle.) It is still extremely tempting to imagine that vision is like television, and that those spike trains get transduced "back into subjective color and sound" and so forth, but we know better, don't we? We don't have to strike up the little band in the brain to play the music we hear in our minds, and we don't have to waft molecules through the cortex to be the grounds for our savoring the aroma of bacon or strawberries. There is no second transduction. And if there were, there would have to be a third transduction, back into spike trains, to account for our ability to judge and act on the basis of our subjective experiences. There might have been such triple transductions, and then there would have been a Cartesian Theater Deluxe, like the wonderful control room in the film Men in Black. But biology has been thrifty in us: it's all done through the medium of spike trains in neurons. (I recognize that dualists of various stripes—a genus thought extinct not so many years ago—will want to dig in their heels right here. I will ignore their howls for the time being, thinking that I can dispatch them later in the argument when I provide an answer to their implied question "What else could it be?")

So there is no *ME*dium into which spike trains are transduced. Spike trains are discriminated, elaborated, processed, reverberated, reentered, combined, compared, and contrasted—but not transduced into anything else until some of them activate effectors (neuromuscular

junctions, hormone releasers, and the like) which do the physical work of guiding the body through life. The rich and complex interplay between neurons, hundreds of neuromodulators, and hormones is now recognized, thanks to the persuasive work of Damasio and many others, as a central feature of cognition and not just bodily control, and one can speak of these interactions as transduction back and forth between different media (voltage differences and biochemical accumulations, for instance)—but none of these is the imagined *ME*dium of subjective experience.

So there just is no home in the brain for qualia as traditionally conceived. My point can be clarified by a simple comparison between two well-understood media: cinema film and digital media. First imagine showing some stone-age hunter-gatherers a movie using a portable Super-8 film projector. Amazing, they would think, but when they were then shown the frames of film up close, they would readily understand—I daresay—that this was not magic, because there were little blobs of color on each frame. (The soundtrack might still be baffling, but perhaps they would hold the film up to their ears and decide, eventually, that the sounds were just too faint for them to hear with their naked ears.) Then show them a film on a portable DVD player, and demonstrate the powers of the removable, interchangeable disks, and let them ponder the question of how such a disk managed to store all the sounds and colors they just observed on the screen. It would probably be tempting for them to declare that it *must* be magic—dualism, in other words. But with a little instruction, they could no doubt catch on to the idea that you don't have to represent color with color, sound with sound. You can *transduce* color, sound—anything, really—into a system of patterns of differences (0s and 1s, spike trains, …) and then *transduce* the elements of that system back into color and sound with playback equipment. This could lay magic to rest.

I had better make my implicit claim explicit, at the risk of insulting some readers: if you think there *has* to be a medium in the brain (or in a dualistic mind) in which subjective colors, sounds, and aromas are *rendered*, you are making the stone-ager mistake. This, I have come to believe, is the stone wall separating my view from wider acceptance. People pay attention to my arguments, and then, confronted with the prospect that qualia, as traditionally conceived, are not needed to explain their subjectivity, they just dismiss the idea as extravagant. "OF COURSE there are qualia!" This thought experiment is meant to shock them: your confidence here, I am saying, is no better grounded than the imagined confidence of the stone-agers that there just *have to be* colors and sounds on the DVD for it to convey colors and sounds to the playback machine. A failure of imagination mistaken for an insight into necessity. "But when I have a tune running through my head, it has pitch and tempo, and the timbre of the instruments is there just as if I were listening to a live performance!" Yes, and for that to be non-magically the case, there has to be a representation of the tune that progresses more or less in real time, and that specifies pitch and timbre, but that can all be accomplished without transduction, without further *rendering*, in the sequence of states of neural excitation in auditory cortex.

Vision isn't television, and audition isn't radio. We are accustomed, now, to playback devices that do transduce the signals back into the colors and sounds from which they were transduced, but we need to take advantage of our twenty-first century sophistication and recognize that the second transduction is optional! The information is in the signal, and all that information can be processed, discriminated, translated, re-coded, simplified, embellished, categorized, tagged, adjusted, and used to guide behavior without ever being transduced back into colors and sounds (or "subjective" colors and sounds).

3 It still seems that qualia exist

But it sure seems that qualia exist, in spite of the foregoing! How could they not? Aren't they needed, for instance, to be the source or cause of our judgments about them? If I have a conviction that I'm seeing an American flag after-

image (see figure 1), and note that the lowest short red stripe intersects the central cross, doesn't there have to be the red stripe I deem myself to be experiencing? Isn't the presence of that red stripe *somewhere* a necessary condition for me seeming to see a red stripe? No, and the alternative has been at least dimly understood since Hume's brilliant discussion of our experience of causation.

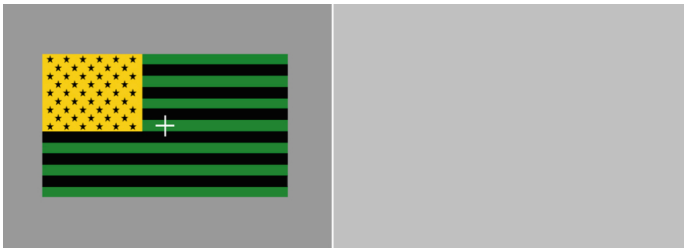

Figure 1: Inverted American Flag.

Consider what I will call Hume's Strange Inversion (cf. Dennett 2009). We think we see causation because the causation in the world directly causes us to see it—the same way round things in daylight cause us to see round things, and tigers in moonlight cause us to see tigers. When we see the thrown ball causing the window to break, the causation itself is somehow perceptible "out there." Not so, says Hume (1739, section 7 "Of the idea of necessary connexion"). What causes us to have the idea of causation is not something external but something internal. We have seen many instances of As *followed by* Bs, Hume asserts, and by a process of roughly Pavlovian conditioning (to put it anachronistically) we have been caused by this series of experiences to have in our minds a disposition, when seeing an A, to expect a B—even before the B shows up. When it does, this *felt* disposition to expect a B is mis-identified as an external, *seen* property of causation. We think we experience causation between A and B, when we are actually experiencing our internal judgment "here comes a B" and "projecting" it into the world. This is a special case of the mind's "great propensity to spread itself on external objects" (Hume 1739, I, xiv). In fact, Hume insisted, what we do is misinterpret an inner "feeling"—an anticipation—as an external property. The "customary transition" in our minds is the source of our sense of causation, a quality of "perceptions, not of objects," but we mis-attribute it to the objects, a sort of benign user-illusion, to speak anachronistically again. As Hume notes, "the contrary notion is so riveted in the mind" that it is hard to dislodge. It survives to this day in the typically unexamined assumption that all perceptual representations must be flowing inbound from outside.

> Hume wrote that the 'mind has a great propensity to spread itself on external objects' (T 1.3.14.25; SBN 167) and that we 'gild and stain' natural objects 'with the colours borrowed from internal sentiment' (EPM Appendix 1.19; SBN 294). These metaphors have invited a further one: that of 'projection' and its cognates. Though not Hume's own, the projection metaphor is now so closely associated with him, both in exegetical and non-exegetical contexts, that the phrase 'Humean projection' is something of a cliché in philosophical discourse. (Kail 2007, p. 20)

Here are a few other folk convictions that need Strange Inversions: sweetness is an "intrinsic" property of sugar and honey, which causes us to like them; observed intrinsic sexiness is what causes our lust; it was the funniness out there in the joke that caused us to laugh (Hurley et al. 2011). There is no more familiar and appealing verb than "project" to describe this effect, but of course everybody knows it is only metaphorical; colors aren't literally projected (as if from a slide projector) out onto the front surfaces of (colorless) objects, any more than the idea of causation is somehow beamed out onto the point of impact between the billiard balls. If we use the shorthand term "projection" here to try to talk, metaphorically, about the mismatch between manifest and scientific image (Sellars 1962), what is the true long story? What is literally going on in the scientific image? A large part of the answer emerges, I propose, from the predictive coding perspective. Every organism, whether a bacterium or a member of *Homo sapiens*, has a set of things in the world that matter to it and which it (therefore) needs to

discriminate and anticipate as best it can. Call this the ontology of the organism, or the organism's "Umwelt" (von Uexküll 1957). This does not yet have anything to do with consciousness but is rather an "engineering" concept, like the ontology of a bank of elevators in a skyscraper: all the kinds of things and situations the elevators need to distinguish and deal with. An animal's "Umwelt" consists in the first place of affordances (Gibson 1979), things to eat or mate with, openings to walk through or look out of, holes to hide in, things to stand on, and so forth. We may suppose that the "Umwelt" of a starfish or worm or daisy is more like the ontology of the elevator than like our manifest image. What's the difference? What makes our manifest image manifest (to us)?

4 Bayesian expectations

Here is where Bayesian expectations (see Clark 2013) could play an iterated role: our ontology (in the elevator sense) does a close-to-optimal job of representing the things in the world that matter to the behavior our brains have to control (cf. Metzinger 2003, on our world models). Hierarchical Bayesian predictions accomplish this, generating affordances galore: we expect solid objects to have backs that will come into view as we walk around them, doors to open, stairs to afford climbing, cups to hold liquid, etc. But among the things in our Umwelt that matter to our wellbeing are ourselves! We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will expect next! And we do. Here's an example:

Think of the cuteness of babies. It is not, of course, an "intrinsic" property of babies, though it seems to be. What you "project" out onto the baby is in fact your manifold of "felt" dispositions to cuddle, protect, nurture, kiss, coo over, … that little cutie-pie. It's not just that when your cuteness detector (based on facial proportions, etc.) fires, you have urges to nurture and protect; you expect to have those very urges, and that manifold of expectations just is the "projection" onto the baby of the property of cuteness. When we expect to see a baby in the crib, we also expect to "find it cute"—that is, we expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world with which we are interacting has the properties we expected it to have. Without the iterated expectations, cuteness could do its work "subliminally," outside our notice; it could be part of our "elevator ontology" (the ontology that theorists need to posit to account for our various dispositions and talents) but not part of *our* ontology, the things and properties we can ostend, reflect on, report, discuss, or appeal to when explaining our own behavior (to ourselves or others). Cuteness as a property passes the Bayesian test for being an objective structural part of the world we live in (our *manifest* manifest image), and that is all that needs to happen. *Any further "projection" process would be redundant. What it is* to experience a baby as cute is to generate the series of expectations and confirmations just described. What is special about properties like sweetness and cuteness is that their perception depends on particularities of the nervous systems that have evolved to make much of them. The same is of course also true of colors. This is what is left of Locke's (and Boyle's) distinction between primary and secondary qualities.[1]

Similarly, when we feel the urge to judge something about "that red stripe" (in the American flag afterimage (see Figure 1) that hovers in our visual field, we have the temptation to insist that there is a red stripe—there has to be!—causing us to seem to see it. But however natural and human this temptation is, it must be resisted. We can be caused to seem to see something by something that shares no features with the illusory object. (Remember Ebenezer Scrooge saying to Marley's ghost: "You may be an undigested bit of beef, a blot of mustard, a crumb of cheese, a fragment of an underdone potato. There's more of gravy than of grave about you, whatever you are!" Many would insist that there has to be a ghost-shaped intermediary in the causal chain between blot of

1 The material in the previous five paragraphs is adapted from Dennett (2013).

mustard and belief in Marley, but Scrooge might be right in addressing his remark to the cause of his current condition, and be leaving nothing Marley-shaped out.) And as for the idea that without being *rendered* such contents are causally impotent, it is simply mistaken, as a thought experiment will reveal. Suppose we have a drone aircraft hunting for targets to shoot at, and suppose that the drone is equipped with a safety device that is constantly on the lookout for red crosses on buildings or vehicles—we don't want it shooting at ambulances or field hospitals! With its video eye it takes in and transduces (into digital bit streams) thirty frames a second (let's suppose) and scans each frame for a red cross (among other things). Does it have to project the frame onto a screen, transducing bit streams into colored pixels? Of course not. It can make judgments based on un-transduced information—in fact, it can't make judgments based on anything else. Similarly your brain can make judgments to the effect that there is a red stripe out there on the basis of spike train patterns in your cortex, and then act on that judgment (by causing the subject to declare "I seem to see a red stripe," or by adjusting an internal inventory of things in the neighborhood, or …). (I am deliberately using the word "judgment" for the drone's discriminations and the brain's discriminations; I have elsewhere called such items micro-takings or content-fixations. The main point of using "judgment" is to drive home the claim that these events are *not* anything like the exemplification of properties, intrinsic or otherwise. They are not qualia, in other words. Qualia—as typically conceived—would only get in the way. Don't put a weighty LED pixel screen in a drone if you want it to detect red crosses, and don't bother installing qualia in a brain if you want it to have color vision. Whatever they are, qualia are unnecessary and may be jettisoned without loss.)

So the familiar idea (familiar in the context of Block's proposed distinction between access consciousness and phenomenal consciousness) that phenomenal consciousness (= qualia) is the basis for access consciousness (= judgments about qualia, qualia-guided decisions, etc.) is backwards.[2] Once the discerning has happened in the untransduced world of spike trains, it can yield a sort of Humean projection —of a red stripe or red cross or just red, for instance—into "subjective space."

But what is this subjective space in which the projection happens? Nothing. It is a theorist's fiction. The phenomenon of "color phi" nicely illustrates the point. When shown, say, two disks displaced somewhat from each other, one sees the apparent motion of a single disk—the phi phenomenon that is the basis of animation (and motion pictures in general). If the disks are of different colors—the left disk red and the right disk green, for instance—one will see the red disk moving rightward and changing its color to green in mid-trajectory. How did the brain "know" to move the disk rightward and switch colors before having access to the green disk at its location? It couldn't (supposing precognition to be ruled out). But it could have Bayesian expectations of continuous motion from place to place that provoke a (retrospective) expectation of the intermediate content, and this expectation encounters no disconfirmation (if the timing is right), which suffices to establish in reality the illusory sequence in the subject's manifest image. So the visual system's *access* to the information about the green disk is causally prior to the "*phenomenal*" motion and color change. Here is a diagram of color phi

2 I once had an occasion to point out this prospect to Block. He had just participated in a laterality test, to see how strongly lateralized for language his brain was. There are two oft-used ways of testing this: with dichotic headphones, which send different words into each ear, where the subject is asked to identify the word heard (typically you only hear one of them!). A second, visual test involves looking at a center target on a screen and having a word or non-word (e.g., "flum" or "janglet") flashed briefly in either the left or right visual field. The subject presses the word button or the non-word button and latencies and errors are recorded. If you are strongly lateralized left (your left hemisphere is strongly dominant for language and does most of the work of language processing), you are faster and more accurate on words and nonwords flashed to the right hemifield. Ned had taken the visual test, and I asked him what he had learned. He was, he said, strongly lateralized left for language, like most people, and he added "the words flashed on the left actually seemed blurry!" I asked him whether the words seemed blurry because he noted the difficulty he was having with them, or whether he had the difficulty because the words were blurry. He acknowledged that he had no introspective way of distinguishing these two hypotheses. Supposing that Block doesn't have some remarkable problem with his eyes, in which the left half of each lens is occluded or misshapen, producing a blur on the left side of his retinas, it is highly likely that the blurriness he seemed to experience was an effect of his felt difficulty in responding, not the cause of this difficulty.

from Consciousness Explained (and Dennett & Kinsbourne 1992):

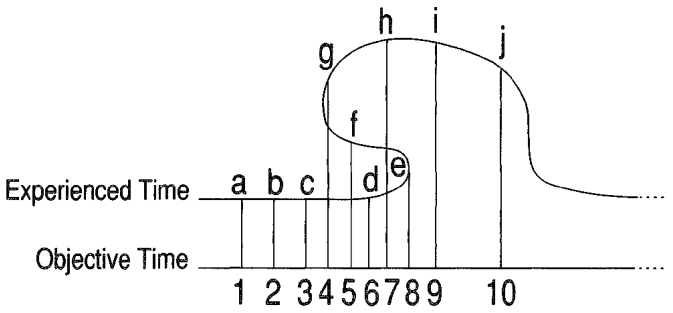

Figure 2: Superimposition of subjective and objective sequences.

In order to explain "temporal anomalies" of conscious experience, we need to appreciate that not only do we not have to represent red with something red, and round with something round; we don't have to represent time with time. Recall my example "Tom arrived at the party after Bill did." When you hear the sentence you learn of Tom's arrival before you learn of Bill's, but what you learn is that Bill arrived earlier. No revolution in physics or metaphysics is needed to account for this simple distinction between the temporal properties of a representation and the temporal properties represented thereby. It is quite possible (in color phi, for instance) for the brain to discern (in objective time) first one red circle (cat time 3) and then a green circle (fat time 5) displaced to the right, and then to (mis-)represent an intermediate red-turning-green circle (eat time 8) yielding the subjective judgment of apparent motion with temporally intermediate color change. Here our Bayesian probabilistic anticipator is caught in the act, jumping to the most likely conclusion in the absence of any evidence. Experienced or subjective time doesn't line up with objective time, and it doesn't have to. The important point to remember from the diagram is that the subjective time sequence is NOT like a bit of kinked film that then has to be run through a projector somewhere so that c is followed by e is followed by f in real time. It is just a theorist's diagram of how subjective time can relate to objective time. Subjective time is not a further real component of the causal picture. No *rendering* is necessary, the judgment is already in, and doesn't have to be re-presented for another act of judging (in the Cartesian Theater).

The temptation to think otherwise may run deep, but it is fairly readily exposed. Consider fiction. Sherlock Holmes and Watson seem real when one is reading a Conan Doyle mystery —as real as Disraeli or Churchill in a biography. When Sherlock seems real, does this require him and his world to be rendered somewhere, in—let's call it—*fictoplasm?* No. There is no need for a medium of fictoplasm to render fiction effective, and there is no need for a mysterious medium, material or immaterial, to render subjective experience effective. No doubt the temptation to posit the existence of fictoplasm derives from our human habit, when reading, of adding details in imagination that aren't strictly in the book. Then, for instance, when we see a film of the novel, we can truly say "That's not how I imagined Holmes when I read the book."

Isn't such rendering in imagination while reading a novel a case of *transduction* of content from one medium (written words as seen on the page) into another (imagined events as seen and heard in the mind's eye and ear)? No, this is not transduction; it is, more properly, a variety of *translation, effortlessly expanding the content thanks to the built-in Bayesian prediction mechanisms.* We could construct, for instance, a digital device that takes problems in plane geometry presented in writing ("From Euclidean axioms prove the Pythagorean Theorem.") and solves them through a process that involves making Euclidean constructions, with all the sides and angles properly represented and labeled, and utilizing them in the proof. The whole process from receipt of the problem to delivery of the called-for proof (complete with printed-out diagrams if you like), is conducted in a single medium of digital bit strings, with no transduction until the printer or screen is turned on to render the answer. (A more detailed description of this kind of transformative process without transduction is found in my discussion (1991) of how the robot Shakey discriminated boxes from pyramids.)

Consider Figure 2 above. Does the access/phenomenal consciousness distinction get depicted therein? If so, access consciousness should be identified with the objective time line, and phenomenal consciousness (if it were something real in addition to access consciousness) would be depicted in the line that doubles back in time. The content feature that creates the kink is an effect of a judgment or discernment that came later in objective time than the discernment of the green circle at time 5. It is because the brain already had access to red circle, then green circle that it generated a representation (but not a *rendering*) of the in-between red-turning-green circle as an elaborative effect.[3]

5 Why do qualia seem so simple and ineffable?

Qualia seem atomic to introspection, unanalyzable simples—the smell of violets, the shade of blue, the sound of an oboe—but this is clearly an effect of something like the resolution of our discernment machinery.

If our vision were as poorly spatially resolved as our olfaction, when a bird flew by, the sky would suddenly "go all birdish,"—that peculiar, indescribable birdishness that one would experience in the visual presence of birds. And this resolution is variable: music lovers and wine enthusiasts and others can train up their ear and their palate and come to distinguish, introspectively, the combining elements of what used to seem atomic and unanalyzable. David Huron (2006), has done some ingenious work teasing out and explaining the combinations of neuroarchitectonic properties that explain the otherwise ineffable characteristic qualia of scale tones (the way *do* sounds different from *re* and *mi* and *so*). It turns out that these "qualia" are actually highly structured properties of neural representations. The explanation, needless to say, is ultimately in the medium of spike trains.

But why should the resolution (if that is the right term) be so low? Why should our brains ignore so much detail in the representations to which "we" have "access"? Minsky (1985), Dennett (1991), Norretranders (1999), Metzinger (2003), and others have said that it is the brain's own access to its own complex internal activities that accounts for the simplicity. This is the brain's effective user-illusion for itself, in much the way the desktop with its icons and various metaphors (click and drag, highlighted targets, etc.) is an elegantly designed user-illusion for laypeople who don't need to know how their computers work.

The brain does not have a single internal witness or homunculus, but it does need something like a lingua franca to get the different and semi-independent subsystems to communicate with each other. (For instance, in the Global Neuronal Workplace model[4] of Dehaene et al. (2006), and others, one should not take it for granted that the *local* meanings of spike train patterns—in the dorsal vision stream, say, or the olfactory bulb—are readily "understood" by all the elements to which some of these signals are broadcast.) I think there is bound to be some important truth in that theme, but it is only part of the story.

6 Whose access?

I think the more interesting suggestion is that the effective "we" when we talk about what "we" have access to, is, indeed, *we*—not just *I*, but *you and me.* It is, more particularly, *your* access to *my* mind that simplifies the information that *we* have access to!

The linguist Stephen Levinson (2006) has studied the remarkable language, Yélî Dnye, of the three thousand or so inhabitants of Rossel Island in the South Pacific—to the north of Papua New Guinea. It is a completely isolated language, unlike any other in the world in many regards. In particular, it is hideously complex, with:

> the largest phoneme inventory (ninety distinct segments) in the Pacific, and many

3 Thanks to David Gottlieb for drawing my attention to this way of looking at access consciousness.

4 Isn't the Global Neuronal Workplace the derided Cartesian Theater after all? No, because what goes on there is not transduction-and-rendering, but informational integration: the coalition and consilience of competing elements. There is no transduction threshold that determines the time-of-entry "into consciousness", and none of the multiple drafts competing in it are singled out as being conscious except retrospectively. This is the point of my admonition always to ask the Hard Question: "And then what happens?" (Dennett 1991, p. 225)

> sounds (such as doubly articulated labial coronal stops) that are either unique or rare in the languages of the world. Among the fifty-six consonants are many multiply articulated segments: e.g., /ʇpn̪m/ is a single segment made by simultaneously putting the tongue behind the alveolar ridge, trilling the lips , and snorting air through the nose. [...] Once the learner is past the sound hurdle, he or she faces another formidable obstacle. The language has an extremely complex system of verb inflection (with thousands of distinct inflectional forms). [...] In addition, substitute forms are used where the subject has been mentioned before, is close or visible, is in motion, or where the sentence is counterfactual or negative, thus providing well over a thousand possibilities [...]. (Levinson 2006, p. 20)

Levinson reports, not surprisingly, that "[h]ardly any mature individuals (such as non-native spouses) who have immigrated into the island community ever learn to speak the language, and children of expatriate Rossels do not fully acquire it from their parents alone." His explanation is speculative, but plausible: a language, left to itself for centuries, will grow ever more complex, like an unpruned bush, simply because it can. The extreme isolation of Rossel Island over the centuries (for various geographic reasons) means that the language has hardly ever been confronted with non-native speakers of another language with whom communication is imperative, for one reason or another. The need for communication soon generates a small cadre of bi-lingual interpreters, and maybe also a pidgin (and maybe later a creole), and all of these alien interfaces work to simplify a language. The least learnable, most baroque (in the sense of exceeding the functional) features of the language are dropped under this pressure. We can see it happening with English today, with simplified dialects such as Emblish (as spoken at the European Molecular Biology Laboratory in Heidelberg) arising naturally and imperceptibly.

I would like to speculate that a similar process of gradual but incessant simplification has shaped the language we have available to explain and describe our minds to each other. Wittgenstein's famous claim about the impossibility of a private language has not weathered the storms of controversy particularly well, but there are neighboring claims—empirical claims—that deserve consideration. Many years ago, Nicholas Humphrey (1987) made the point that has begun to attract adherents today:

> While it is of no interest to a person to have the same kind of kidney as another person, it is of interest to him to have the same kind of mind: otherwise as a natural psychologist he'd be in trouble. Kidney transplants occur very rarely in nature, but something very much like mind-transplants occur all the time [...]. [So] we can assume that throughout a long history of evolution all sorts of different ways of describing the brain's activity have been experimented with but only those most suited to doing psychology have been preserved. Thus the particular picture of our inner selves that human beings do in fact now have—the picture we know as 'us', and cannot imagine being of any different kind—is neither a necessary description nor is it any old description of the brain: it is the one that has proved most suited to our needs as social beings. That is why it works. Not only can we count on other people's brains being very much like ours, we can count on the picture we each have of what it's like to have a brain being tailor-made to explain the way that other people actually behave. Consciousness is a socio-biological product—in the best sense of socio and biological. (p. 18)

Chris Frith, for instance, has recently taken up the theme (in conversation) that consciousness has some features, because everything in consciousness has to be couched in terms that can be communicated to other people readily.

The ineffability barrier we all experience when trying to tell others what it is like to be

us on particular occasions is highly variable, not just between individuals, but over time within a single individual, as a result of formal or informal training. It plays a dynamic role in shaping the contents of our consciousness over time.[5] (This would be true only for human consciousness, obviously.)

7 A thought experiment: Mr. Capgras

Finally, it might seem that whereas some subjective properties—cute, sweet, funny, sexy, the characteristic sounds of scale tones—might be accounted for in terms of Bayesian expectations about how one will be disposed to behave in their presence, the very simplicity of colors must block any attempt to treat them in a similar fashion. There is no way one expects to behave in the presence of navy blue, or pale yellow, or lime green. So it may seem, but this is itself an artifact of our penchant for thinking—as Hume famously did—of colors as simples. Hume was discountenanced by the notorious missing shade of blue, and found it ideologically inconvenient to suppose, as we now know, that color experience is in fact highly complex and compositional, and deeply anchored in dispositions of our perceptual systems.[6] Moreover, color experiences are no more atomic than scale tone experiences, and give rise to all manner of expectations, which tend to go unnoticed, but can be thrown into sharp focus by a thought experiment: my fantasy about poor Mr. Clapgras, the man who wakes up to find all his emotional dispositions with regard to colors inverted while leaving intact his cognitive habits and powers (see Dennett 2005, pp. 91–102, for a more detailed account, with objections considered and rebutted). Ex hypothesi, Mr. Clapgras identifies colors and sorts colors correctly (he does not suffer from the well-studied conditions color anomia, or cerebral achromatopsia), but he finds the world disgusting, unbearable. Food looks just terrible to him now, and he has to eat blindfolded, since his emotional responses to all colors have shifted 180 degrees around the color circle (Grush this collection). He calls shocking pink "shocking pink" but marvels at the inappropriateness of its name. The only way we can explain his distress is by observing that he notices that something is wrong—which has to mean he was expecting something else. He is surprised that breaking a fresh egg into a frying pan on a sunny morning doesn't bring a smile to his face, that a glimpse of his obnoxious neighbor's lime green convertible doesn't irritate him the way it used to do, that he feels no stirring of childhood patriotism when he sees the red white and blue waving in the breeze. Like the sufferers of Capgras delusion, poor Mr Clapgras senses a disturbance: something is very wrong, but it isn't the evaporation of intrinsic internal properties.

8 Conclusion

The considerations I have raised in this essay are not new, but perhaps bringing them together as I have done will help show that a counter-intuitive theory like mine still has an advantage over some of the fantasies in which philosophers have recently indulged. It may well be, as Paul Bloom (2004) has suggested, that we are all "natural born dualists," but just as eyeglasses can correct for myopia, natural-born or not, so science can correct for this innate cognitive disability. Intuitions to the contrary are important data, but should not be taken to indicate a limitation of science, as some have thought. In fact, if the best scientific theory of consciousness turns out not to be deeply counterintuitive at first, among the data it will have had to explain is why it took us so long to arrive at it.

Acknowledgements

Thanks to Michael Cohen, and to Andy Clark, and the rest of Dmitry Volkoff's Greenland Consciousness and Free Will workshop (June 10-17, 2014), for editorial advice on this essay.

5 Note that I am not saying that our day-to-day consciousness wouldn't occur in the absence of human company, but an implication of my speculation is that a Robinson Crusoe human, somehow raised from birth without human contact, would have subjectivity more inaccessible to us—once we discovered him and attempted to communicate with him—than the speech acts of the Rossel Islanders.

6 In Cohen & Dennett, we point out that limbic or emotional responses to colors have to count as instances of "access" to color-representing states "however coarse-grained or incomplete, because such a reaction can obviously affect decision making or motivation" (2011, p. 5).

References

Bloom, P. (2004). *Descartes' baby: How the science of child development explains what makes us human.* New York, NY: Basic Books.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-253. 10.1017/S0140525X12000477

Cohen, M. A. & Dennett, D. C. (2011). Consciousness cannot be separated from function. *Trends in Cognitive Sciences*, *15* (8), 358-365. 10.1016/j.tics.2011.06.008

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J. & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences*, *10* (5), 204-211. 10.1016/j.tics.2006.03.007

Dennett, D. C. (1978). "What's the difference: Some riddles," (commentary on Puccetti and Dykes). *Behavioral and Brain Sciences*, *1* (3), 351-351.

——— (1991). *Consciousness explained.* Boston, MA: Little, Brown and Company.

——— (1994). Get real. *Philosophical Topics*, *22* (1-2), 505-568.

——— (1995). "The path not taken," commentary on Ned Block, "On confusion about a function of consciousness". *Behavioral and Brain Sciences*, *18* (2), 252-253.

——— (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness.* Cambridge, MA: MIT Press.

——— (2009). Darwin's 'strange inversion of reasoning'. *Proceedings of the National Academy of Sciences*, *106* (1), 10061-10065. 10.1073/pnas.0904433106

——— (2013). Expecting ourselves to expect: The Bayesian brain as a projector (commentary on Clark, 2013). *Behavioral and Brain Sciences*, *36* (3), 209-210. 10.1017/S0140525X12002208

Dennett, D. C. & Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Behavioral and Brain Sciences*, *15* (2), 183-247. 10.1017/S0140525X00068229

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston, MA: Houghton Mifflin.

Grush, R., Jaswal, L., Knoepfler, J. & Brovold, A. (2015). Visual Adaptation to a Remapped Spectrum: Lessons for Enactive Theories of Color Perception and Constancy, the Effect of Color on Aesthetic Judgments, and the Memory Color Effect. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Hume, D. (1739). *Treatise of human nature.* London, UK: John Noon.

Humphrey, N. (1987). *"The uses of consciousness" The 57th James Arthur Lecture.* New York, NY: American Museum of Natural History.

——— (2006). *Seeing red: A study in consciousness.* Cambridge, MA: Harvard University Press.

——— (2011). *Soul dust: The magic of consciousness.* Princeton, NJ: Princeton Universiry Press.

Hurley, M., Dennett, D. C. & Adams , jr. , R. B. (2011). *Inside jokes: Using humor to reverse-engineer the mind.* Cambridge, MA: MIT Press.

Huron, D. (2006). *Sweet anticipation; Music and the psychology of expectation.* Cambridge, MA: MIT Press.

Kail, P. J. E. (2007). *Projection and realism in Hume's philosophy.* Oxford, UK: Oxford University Press.

Levinson, S. C. (2006). Introduction. In S. C. Levinson & P. Jaisson (Eds.) *Evolution and culture* (pp. 1-42). Cambridge, MA: MIT Press.

Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

——— (2009). *The ego tunnel. The science of the mind and the myth of the self.* New York, NY: Basic Books.

Minsky, M. (1985). *The society of minds.* New York, NY: Simon & Schuster.

Norretranders, T. (1999). *The user illusion: Cutting consciousness down to size.* London, UK: Penguin Press Science.

Prinz, J. (2012). *The conscious brain.* Oxford, UK: Oxford University Press.

Sellars, W. (1962). Philosophy and the scientific image of man. In R. Colodny (Ed.) *Frontiers of science and philosophy* (pp. 35-78). Pittsburgh, PA: University of Pittsburgh Press.

von Uexküll, J. (1957). A stroll through the worlds of animals and men: A picture book of invisible worlds. In C. H. Schiller (Ed.) *Instinctive behavior: The development of a modern concept.* New York, NY: International Universities Press.

Qualia explained away

A Commentary on Daniel C. Dennett

David H. Baßler

In his paper "Why and how does consciousness seem the way it seems?", Daniel Dennett argues that philosophers and scientists should abandon Ned Block's distinction between access consciousness and phenomenal consciousness. First he lays out why the assumption of phenomenal consciousness as a second medium is not a reasonable idea. In a second step he shows why beings like us must be convinced that there are qualia, that is, why we have the strong temptation to believe in their existence. This commentary is exclusively concerned with this second part of the target paper. In particular, I offer a more detailed picture, guided by five questions that are not addressed by Dennett. My proposal, however, still resides within the framework of Dennett's philosophy in general. In particular I use the notion of intentional systems of different orders to fill in some details. I tell the counterfactual story of some first-order intentional systems evolving to become believers in qualia as building blocks of their world.

Keywords
Dispositions | Intentional systems | Predictive processing | Qualia | Zombic hunch

Commentator

David H. Baßler
davidhbassler@gmail.com
Johannes Gutenberg-Universität
Mainz, Germany

Target Author

Daniel C. Dennett
daniel.dennett@tufts.edu
Tufts University
Medford, MA, U.S.A.

1 Introduction

The first of Rapoport's Rules[1] for composing a critical commentary states that one should present the target view in the most charitable way possible (Dennett 2013a). Although I generally agree with many of Daniel Dennett's views, especially his argument against the existence of qualia (constituting the first part of the target paper), the diagnosis that there is the *zombic hunch*,[2] along with his strategy for explaining why it exists, the connection between qualia and predicted dispositions, was hard to grasp. Dennett presents the idea that when we talk about qualia, what we really refer to are our dispositions in earlier works (e.g., Dennett 1991). But the connection to predictive pro-

1 Dennett named these rules after social psychologist and game theorist Anatol Rapaport. They are not to be confused with another "Rapoport's Rule", named after Eduardo H. Rapaport (cf. Stevens 1989). Here is the full list of Dennett's Rapaport's Rules:
1. "You should attempt to re-express your target's position so clearly, vividly, and fairly that your target says, 'Thanks, I wish I'd thought of putting it that way.'"
2. "You should list any points of agreement (especially if they are not matters of general or widespread agreement)."
3. "You should mention anything you have learned from your target."
4. "Only then are you permitted to say so much as a word of rebuttal or criticism."
(Dennett 2013a, p. 33)

2 A philosophical zombie has nothing to do with any other sort of zombie. It behaves in *every* way like a normal person. The only difference is, that it lacks phenomenal experiences (though *ex hypothesi* it believes that is has phenomenal experiences). The zombic hunch is the intuition that a philosophical zombie would be different from us.

cessing is new (see also Dennett 2013b). There still seem to be some stepping stones missing, which I hope to fill in with my reconstruction. My goal is to provide a complete story that sticks as close to Dennett's argument as possible. This paper is not supposed to be a "rebuttal" or "criticism", but an "attempt to re-express [Dennett]'s position" (see footnote 1).

The structure of this commentary is as follows: in the first section I shall give a short outline of Dennett's explanation of why we have the zombic hunch. Since this involves the predictive processing framework, I shall give a very short introduction to this first. Following this, I present a short list of five questions that have not, in my opinion, yet been sufficiently addressed. In the second section I present an interpretation, or perhaps an extension, of Dennett's answers to these questions, by relying on the concept of an intentional system and using a strategy involving telling the counterfactual story of the evolution of some agents who end up believing in qualia (although *ex hypothesi* there are none). In the third section I shall analyze which features qualia should have, according to the beliefs of these agents, and show that there is at least a significant overlap with features many consider qualia to have.

I want to give a short justification for the unorthodox way of accounting for *beliefs* about x instead of for x's existence itself. This is a general strategy found in other areas of Dennett's work. For example, he has asked, "Why should we think there is intentionality although there is none?" (Dennett 1971), "Why should we believe there is a god although there is none?" (Dennett 2006), and "Why should we think there is a problem with determinism and free will although there is none?" (Dennett 1984, 2004). Dennett's philosophy can in parts be seen as a therapeutic approach to "philosopher's syndrome"—"mistaking failures of imagination for insights into necessity" (e.g., Dennett 1991, p. 401; Dennett 1998a, p. 366)—by making it easier to see why we are convinced of the existence of something, even when there are good reasons to believe that it doesn't exist.

I want to draw attention to Hume's *Of Miracles* (Hume 1995, X), where he states that the likelihood of a testimony about miracles being wrong is always greater than the likelihood of the miracle itself. This serves as a nice analogy for the case at hand: we might think of our own mind as a good "witness", but we already know too much about its shortcomings. So we should be suspicious when it cries out for a revolution in science or metaphysics, because this cry rests on the belief that something is missing, when no data but this very belief itself makes the demand necessary. Instead we should examine what else could have led our minds to form this conviction.

2 Dennett's proposal

In "Why and how does consciousness seem the way it seems?" Dennett gives an argument for why philosophers and scientists should abandon Ned Block's distinction between access consciousness and phenomenal consciousness, zombies, and qualia altogether. The argument is twofold: first Dennett lays down his argument for why the assumption of phenomenal consciousness as a second medium whose states are conscious experiences or qualia is "scientifically insupportable and deeply misleading" (Dennett this collection, section 2). It is insupportable because there is simply no need to posit such entities to explain any of our behavior, so for reasons of parsimony they should not be a part of scientific theories (see also Dennett 1991, p. 134). The assumption is deeply misleading because it makes us look for the wrong things, namely, the objects our judgments are about, rather than the causes of these judgments, which are nothing like these objects.

In a second step Dennett shows why creatures like us must be convinced that there are qualia, that is, why we have such a strong temptation to believe in their existence, *even though* there are no good reasons for this (Dennett this collection, section 2 and 3; other places where Dennett acknowledges this conviction, the zombic hunch, are Dennett 1999; Dennett

2005, Ch. 1; Dennett 2013a, p. 283). The following sections are exclusively concerned with this part of the target paper.

After completing the second step, Dennett explains why we ascribe qualia their characteristic properties—simplicity and ineffability (Dennett this collection, section 4 & 5). Although I also say something about this point (see section 4), Section 6 is an intuition pump (cf. e.g., Dennett 2013a) that will help the reader to apply Dennett's alternative view to the experience of colors.

Before I present a short outline of Dennett's second step, I want to briefly describe the predictive processing framework. This is necessary since both Dennett's argument as well as my reconstruction make use of this framework. I shall not go into details of hierarchical predictive processing (PP) accounts here, since at least three papers in this collection (Clark, Hohwy, and Seth), as well as the associated commentaries (Madary, Harkness, and Wiese), are concerned with this topic and also offer ample references for introductory as well as further reading. I will instead give a very short description of the points that are most relevant to Dennett's argument and recommend the above-mentioned papers and the references given there to the interested reader.

2.1 Predictive processing

In the PP framework, the brain refines an internal generative stochastic model of the world by continuously comparing sensory input (extero- as well as interoceptive) with predictions continuously created by the model. The overall model is spread across a hierarchy of layers, where the sensory layer is the lowest and each layer tries to predict (that is, to suppress) the activation pattern of the layer beneath it. The whole top-down activation pattern might be interpreted as a global hypothesis about the hidden causes of ongoing sensory stimulation. The difference between predicted and actual activation (*prediction error*) is what gets propagated up the hierarchy and leads to changes in the hypothesis. To be exact, this is only one possibility. Another is that this leads to an action that changes the input in such a way that the prediction is vindicated (*active* inference, see e.g., Friston et al. 2011). However, although this aspect of PP —that it provides one formally-unified approach to perception and action—is a strength of the framework, it is not important here, given the context of this commentary. These changes are supposed to follow Bayes' Theorem, which is why one might speak of Bayesian prediction (cf. e.g., Hohwy 2013).

The higher the layer in the hierarchy the more abstract the contents and the longer the time-scales or the predictive horizon. One example of a very abstract content is "only one object can exist in the same place at the same time" (Hohwy et al. 2008, p. 691, quoted after Clark 2013, p. 5).

One point to keep in mind is that, according to Hohwy (2014), this framework implies a clear-cut distinction between the mind and the world. That is, there is an *evidentiary boundary* between "where the prediction error minimization occurs" and "hidden causes [of the sensory stimulation pattern] on the other side" (Hohwy 2014, p. 7). I will come back to this point later in this commentary.

2.2 The outline of Dennett's argument

1. Our own dispositions, expectations, etc. are part of the generative self-model instantiated by our brains. "We ought to have good Bayesian expectations about what we will do next, what we will think next, and what we will expect next" (Dennett this collection, p. 5)
2. When our brains do their job (described in (1)) correctly, i.e., there are no prediction-error signals, we misidentify dispositions of the organism with properties of another object. For instance, instead of attributing the disposition to cuddle a baby correctly to the organism having the disposition, our brain attributes "cuteness" to the baby.[3] Color qualia

3 "Think of the cuteness of babies. It is not, of course, an 'intrinsic' property of babies, though it seems to be. […][W]e expect to expect to feel the urge to cuddle it and so forth. When our expectations are fulfilled, the absence of prediction error signals is interpreted as confirmation that, indeed, the thing in the world with which we are in-

and other types of qualia also belong to this category.[4]
3. This means, under a personal level description, that we believe that there are properties *independent of the observer*, such as the cuteness of babies, the sweetness of apples, or the blueness of the sky, etc.
4. This is why it is so hard for us to doubt that qualia exist in the real world.

The crucial points seem to be (1) and (2). Before I lay out my interpretation I want to highlight some points that are not addressed in Dennett (this collection), but which are crucial if we are to have a complete picture. In the section *Our Bayesian brains*, I present a reconstruction that addresses these issues.

2.3 Five questions

1. **Why do we need to monitor our dispositions?** As noted in Dennett (2010), self-monitoring, in the sense of monitoring of our dispositions, values, etc., isn't needed unless one needs to communicate and to hide and share specific information about oneself at will. In his paper, Dennett does not address this issue, yet presupposes that "among the things in our *Umwelt* that matter to our well-being are ourselves". This is obvious if one reads "ourselves" as the motions of our bodies, but not so obvious if one includes things like "what we will think next, and what we will expect next", as Dennett does (Dennett this collection, p. 5). The next question is concerned with this latter form of self-monitoring:
2. **How is self-monitoring accomplished?** Hohwy (2014) refers to an evidential boundary in the predictive processing framework (see the section 2.1): there is a clear distinction between the mind/brain and the world (of which the body without the brain is a part), whose causal structure is yet to be revealed. Our expectations are part of our mind, which, if talk of the boundary is correct, does not have direct access to its own states *as* its own states—the mind is a black box to itself. So the prediction of its expectations needs to be indirect (just like the predictions of the causes of the sensory stimulation in general), and therefore the question arises how the self-monitoring of the mind is achieved according to Dennett. There is a further concern with self-monitoring, which one might call the "acquisition constraint" (cf. e.g., Metzinger 2003, p. 344):
3. **How did this self-monitoring evolve in a *gradual* fashion?** Large parts of Breaking the Spell are dedicated to making understandable how "belief in belief" could have evolved over the centuries, beginning long before the appearance of any religion. Dennett's goal here is quite similar: the explanation aims to make understandable how we came to believe in qualia, etc. But a step-by-step explanation is missing. I consider this form of the acquisition-constraint one of the most crucial for any satisfying explanation of this sort: each single step has to be understandable as one likely to have happened. One reason for this is that it would support a more fine-grained and mechanistic understanding; another is that it would satisfy the gradualism-constraint of Darwinism, which says that minds (just like anything else) "must have come into existence gradually, by steps that are barely discernible *even in retrospect*" (Dennett 1995, p. 200, emphasis in original).

teracting has the properties we expected it to have" (Dennett this collection, p. 5).

4 The intuition pump of Mr. Clapgras in Dennett's section 6 is there to make the point that colors can be seen as dispositional properties of the organism rather than as properties of perceptual objects, in the same way as cuteness. Whether one is convinced by this or not, the intuitive problem seems to be the same: science tells us there are no properties like cuteness or color, while the zombic hunch tells us that this cannot be true. A more detailed discussion can be found in Dennett (1991, p. 375). I will not go into this here, but for the sake of argument I shall assume that this admittedly counter-intuitive categorization is acceptable. The reader's willingness to accept it might be helped by the following point given by Nicholas Humphrey, which reminds us that although at first thought colors do not *seem* to have action-provoking effects (like cuteness or funniness), after second thought one might think differently:
"As I look around the room I'm working in, man-made colour shouts back at me from every surface: books, cushions, a rug on the floor, a coffee-cup, a box of staples—bright blues, reds, yellows, greens. There is as much colour here as in any tropical forest. Yet while almost every colour in the forest would be meaningful, here in my study almost nothing is. Colour anarchy has taken over." (Humphrey 1983, p. 149; quoted in Dennett 1991, p. 384).

Once we know why and how our brains accomplish the task of monitoring our dispositions and how they came to do so, one might still wonder why (as claimed in point 2, page 3) exactly these abstract properties of the organism would be misidentified as concrete properties of other things:

4. **Why do we misidentify our dispositions?** One of Dennett's central claims is that we misidentify our own dispositions, which leads to belief in qualia.[5] Although misidentification seems to be ubiquitous (see superstition, religion, magic tricks, the rubber hand illusion—Botvinick & Cohen 1998; and even full body illusions—Blanke & Metzinger 2009) it nonetheless requires a special explanation in each case: is this a shortcoming of a system that has no disadvantages, or is it even something that benefits the system in some way (cf. McKay & Dennett 2009)? Keeping this last possibility in mind one might ask:
5. **Why are we so attached to the idea of qualia?** There seems to be something more that leads people to believe in qualia. There is the intuition that without qualia we would be very different—we would be "mere machines", we could not *enjoy* things like a good meal or the smell of the air after it rains (a discussion of this characteristic of beliefs-about-qualia can be found in Dennett 1991, p. 383). Some might go further and say that our whole morality rests on the existence of qualia of pain and suffering (this worry is dealt with in Dennett 1991, p. 449). However, what I am concerned with here is not whether it is true that qualia are the basis of our morality, but why we should think them to be so. From the argument presented by Dennett it is not clear why we are so attached to the idea of qualia. It is not obvious why we do not react as disinterestedly to their denial as we did to the revelation that there is no ether.[6] But, as a matter of fact, we react differently: this is not like when any other entity, posited for theoretical reasons, is shown to not exist; it is as if without qualia we couldn't possibly be *us*.

3 An interpretation

3.1 Intentional Systems Theory

An important part of what follows is Intentional Systems Theory (IST). What is crucial here is that according to IST, all there is to being an agent in the sense of having beliefs and desires upon which to act is to be describable via a certain strategy: the *intentional stance*. The intentional stance is a "theory-neutral way of capturing the cognitive competences of different organisms (or other agents) without committing the investigator to overspecific hypotheses about the internal structures that underlie the competences" (Dennett 2009, p. 344). If one predicts the behavior of an object via the intentional stance, one presupposes that it is optimally designed to achieve certain goals. If there are divergences from the optimal path, one can, in a lot of cases, correct for this by introducing abstract entities or false beliefs. Since there are presumably no 100%-optimally-behaving creatures in the world, every intentional profile (a set of beliefs and desires), generated via adoption of the intentional stance, contains a subset of false beliefs.[7] It seems that humans have a "generative capacity [to find the patterns revealed by taking the intentional stance] that is to some degree innate in normal people" (Dennett 2009, p. 342). I will come back to this point and its connection to PP in the next section.

Let us assume for the sake of argument that IST gives a correct explanation of what it is to be an agent (in the sense of someone who has beliefs and desires and acts according to

5 What qualia are [...] are just those complexes of dispositions. When you say 'This is my quale,' what you are singling out, or referring to, whether you realize it or not, is your idiosyncratic complex of dispositions. You seem to be referring to a private, ineffable something-or-other in your mind's eye, a private shadeshade of homogeneous pink, but this is just how it seems to you, not how it is. (Dennett 1991, p. 389).

6 This property of the beliefs is acknowledged in Dennett (2005), p. 22, fn 18: "[The Zombic Hunch] is visceral in the sense of being almost entirely arational, insensitive to argument or the lack thereof".

7 See Dennett (1987) for an elaborate discussion of the intentional stance and its implications, Dennett 1998b for the ontological status of beliefs and desires, Bechtel (1985) for another interesting interpretation, and Yu & Fuller (1986) for a discussion of the benefits of treating beliefs and desires as abstracta.

them), and that PP allows us to see how an agent can be implemented on the "algorithmic level"(see Dennett's discussion in Dennett 1987, p. 74, where he refers to the IST as a "competence model"). Whenever I say that an agent believes, wants, desires, etc. something I mean it in exactly the sense found in IST.

Intentional systems can be further categorized by looking at the content of their beliefs, e.g., a second-order intentional system is an intentional system that has beliefs and/or desires about beliefs and/or desires, that is, it is itself able to take an intentional stance towards objects (Dennett 1987, p. 243). A first-order intentional system has (or can be described as having) beliefs and desires; a second-order intentional system can ascribe beliefs to others and itself. If something is a second-order intentional system it harbors beliefs such as "Peggy believes that there's cheese in the fridge". But taking the intentional stance towards an object is an ability that comes in *degrees*. I now want to describe what one might call an intentional system of *1.5th order*, an intermediate between first- and second-order intentional systems. This is a system that is not able to ascribe full-fledged desires and beliefs with arbitrary contents to others or itself. We, as intentional systems of high order, have no difficulty in ascribing beliefs and desires with very arbitrary contents, such as "She wants to ride a unicorn and believes that following Pegasus is a good way to achieve that goal". But the content of beliefs and desires that such an intentional system of 1.5th order can ascribe should be constrained in the following way:

1. An intentional system of 1.5th order is able to ascribe desires only in a very particular and concrete manner, i.e., actions that the object in question wants to perform with certain particular existing objects, that the system itself knows about (e.g., the desire to eat the carrot over there), but not goals directed at nonexistent objects, described by sentences like "he wants to build a house", or objects the ascriber itself does not know about.
2. It is only able to ascribe beliefs to others that it holds itself. That means it is able to take the basic intentional stance with the default assumption that the target object in question believes whatever is true (if we assume the ascriber's beliefs are in fact all true), but lacks the ability to correct the ascriptions if it leads to wrong predictions for the behavior of the target. A real-world example can be found in Marticorena et al. (2011): rhesus macaques in a false belief task can correctly predict what a person will do, given that the person knows where the object is hidden and they have seen the person getting to know this. They can also tell when a person doesn't have the right knowledge, but they cannot use this information to make a prediction about where the person will look.

The implementation of such an intermediate between first- and second-order intentional systems can be easily imagined following predictive coding principles, as I will soon show. Following this, I argue that this sets down the basic fundaments for systems evolving from this position to be believers in qualia, etc.

The reason for introducing this idea is that I want to show how, given predictive processing principles and a certain selection pressure, a 1.5th-order-intentional-system might develop from a first-order-intentional-system. In a next step, I will argue that under an altered selection pressure such a system might become a full-fledged nth-order-intentional-system, where n is greater or equal to two. Systems evolving in such a way, as I will describe, are bound to believe in the existence of something like qualia. In some sense this is only a just-so story, but the assumed selection pressures are very plausible, and the empirically-correct answer might not be too far away from this.

3.2 Our Bayesian brains[8]

To see how the pieces fit together imagine the situation of some first-order intentional systems, agents, which are the first of their kind. They act according to their beliefs and desires. They do so because the generative models im-

8 This section takes strong inspiration from Wilfrid Sellars' section "Our Rylean Ancestors" in Sellars (1963, p. 178).

plemented in their brains generate a sufficient number of correct predictions about their environment for them to survive and procreate. They do a fairly good job of avoiding harms and finding food and mates. Since they are first-order intentional systems, the behavior of their conspecifics amounts to unexplained noise to them, because they are unable to predict the patterns of most of their behavior (which is what makes them *merely* first-order intentional systems), though they might well predict their behavior as physical objects, e.g., where someone will land if she falls off a cliff, for instance.

When resources are scarce, this leads to competition between these agents and it becomes an advantage to be able to predict the behavior of one's conspecifics. This behavior is by definition pretty complex (they are intentional systems), but one can get some mileage out of positing the following regularity: some objects in the world have properties that lead to predictable behavior in agents, e.g., if there is an apple tree this will lead to the agents approaching it, if they are sufficiently near, etc., whereas if there is a predator, they will run from it, etc. Their model of the world is populated by properties of items that allow the (arguably rough) predictions of *agent behavior*. One might indeed say that the desires of the agents are *projected* [9] onto the world.[10] Those who acquire this ability are now 1.5th order intentional systems (see above; monkeys and chimpanzees might turn out to be such, see Roskies this collection).[11] However, findings in this area are controversial. See Lurz 2010), since they can predict the behavior of others, given that their behavior is indeed explainable via reference to actually-existing objects, such as apples or potential sexual partners. In addition to these properties, there is a new category of objects in "their world": beings that react to these properties in certain ways.[12]

In a next step we might suppose that a system of communication or signaling evolves (the details are not important), turning our intentional systems of 1.5th order into communicative agents. As communicative beings they have an interest in hiding and revealing their beliefs according to the trustworthiness of others and their motives (cf. Dennett 2010). That is, any of those beings needs to have access to what it itself will do next, so that they can hide or share this information, depending on information about the other. One might think of hiding the information about one's desire to steal some food, and so on.

This is a situation where applying the predictive strategy that was formerly only used to explain the behavior of others to *oneself* becomes an advantage for each of the agents.[13] Agents like this believe in the existence of a special kind of special kind of properties, i.e., they predict their *own* behavior on the basis of generative models that posit such properties: they believe that they approach apples *because* they are *sweet*, cuddle babies *because* they are *cute*, laugh about jokes *because* they are *funny*. Applying the strategy to their own behavior puts them in the same category (according to the generative model) as the others: they are unified objects that react to cer-

9 What I mean by "project" is that instead of positing an inner representation whose content is "I (the system in question) want to eat that apple" and whose function is a desire, along with correct beliefs about the current situation, what is posited is an eat-provocative property of the apple itself. Both theoretical strategies allow for the prediction of the same behavior. The crucial difference is that attributing new properties to objects that are already part of the model is a simpler way of extending the model than positing a complex system of internal states to each agent. Thus it is also more likely to happen. It's definitely much simpler than extending the model to incorporate all the entities that explain the behavior on a functional level (i.e., all the neurons, hormones etc.). It is successful to the same extent the intentional stance is successful, that is, in an arguably noisy way, but still successful enough to gain an advantage (since *ex hypothesi* all the conspecifics are intentional systems).

10 This is very close to Gibson's affordances (e.g., Gibson 1986) in that "values and meanings are external to the perceiver" (p. 127) and in a couple of other respects (ibid.). It is, however, different in that the postulated properties serve to predict the behavior of *others* and not to guide the behavior of the organism itself. For the relation between Gibsonian affordances and predictive processing see e.g., Friston et al. (2012).

11 "[R]ecent work on non-human primate theory of mind suggests that monkeys and chimpanzees have a theory of mind that represents goal states and distinguishes between knowledge and ignorance of other agents (the presence and absence of contentful mental representations), even if it fails to account for misrepresentation." (Roskies this collection, p. 12).

12 The selection of goals and other cognitive capabilities, etc., is all placed outside of the target object (see footnote 9). It will approach the object that has the highest attraction value, given that there is no object with a higher repulsion value, i.e., there is no internal selection process represented *as* internal selection. What makes other agents special objects, in this model, is that they react to properties that no other things react to, not that they have an internal life that is somehow special.

13 Notice that according to PP, there is no shortcut to be taken: the mind is a black box to itself—it has to infer its own properties just as any others.

tain properties, not a bunch of cells trying to live among one another.[14]

The agent-models of these beings might improve by integrating the fact that sometimes it is useful to posit non-existing entities or omit existing entities in order to predict the behavior of a given conspecific (think of subjects in the false belief-task looking in the wrong box). By this the concept of (false) beliefs arises. One can imagine how they further evolve into full-fledged second and higher-order intentional systems, in an arms-race for predicting their fellows.[15]

A further step: they develop sciences like we did and will come to have a scientific image of the world, which contains no special simple properties of objects that cause "agents" to behave in certain ways. They come to the conclusion that the brain does its job without taking notice of properties like cuteness or redness, "instead relying" on computations, which take place in the medium of spike trains and nothing but spike trains (cf. target, section 1). Their everyday predictions of others and most importantly of themselves still rely on the posited properties. And some might wonder whether there isn't something missing from the scientific image.

According to the scientific image, they, as biological organisms, react to photons, waves of air, etc., but these are not the contents of their own internal models employed in solving the continuous task of predicting themselves. The simplest things they react to seem to be colors and shapes, (perceived) sounds, etc. The reaction towards babies is explained via facial proportions and the like, but this is far from what their generative models "say", which is "the reaction to babies is caused by their cuteness".

They begin to build robots, which react to babies like they do. They say things like, "all this robot reacts to are the patterns in the baby's face, the proportions one can measure; but although it reacts like we do, it does not do so because of the baby's cuteness". Of course only non-philosophers might say that science misses a property of the baby, but philosophers still see that there is *something* missing, and since cuteness is not a property of the outside world, they conclude that it must be a property of the agents themselves.

This seems to me to be the current situation. We have the zombic hunch because it seems to us that there is something missing and it seems so because our generative models are built upon the assumption that there are properties of things out there in the world to which systems like us react in certain ways. We never consider others like us to be zombies because they are agents like us or better: we are systems like them. We dismiss robots because we know they can only react to measurable properties, which do not *seem* to us to be the direct cause of *our* behavior.

4 An analysis

Is it true that properties such as cuteness do not correspond to anything? In a sense it is false to deny that any such correspondence exists: such properties do correspond to the cuddle-provocativeness of a baby, the eating-provocativeness of an apple, etc., *as a cause of the behavior of agents*. They are "lovely" properties (Dennett 1991, p. 379), and there is a way to measure them: we can use ourselves as detectors. But the reason we, intuitively, do not accept a robot as a subject like ourselves is because we know how the robot does it: we know that it calculates, maybe even in a PP-manner —we know that it does not react directly to the properties that seem to exist and that seem to count. Neither do we, or the beings described above. But their own prediction of themselves treats such complex properties as simple, because there is nothing to be gained by being more precise than is necessary for *sufficiently* accurate prediction.[16]

This is my reconstruction of Dennett's claim that the mind projects its dispositions

14 This is where one might speak of the origin of a self-model (Metzinger 2003) in some sense, where there is not only a model of the body (built up by proprioceptive inputs) but also a model of the self as having (primitive) goals, at least in any given moment.

15 Maybe language plays an important part in this further development as an external scaffold (cf. Clark 1996; Dennett 1994). One fact supporting this view is that monkeys do not seem to be able to understand the concept of false belief (and therefore the concept of belief) (cf. Marticorena et al. 2011, but also Lurz 2010 for an overview of this debate).

16 This is also true of affordances (see e.g., Gibson 1986, p. 141).

onto the world via Bayesian prediction. I want to draw attention to some of the features ascribed to those properties that this story predicts:

1. **These properties are "given directly" to a person**
   The overall generative model depicts the whole organism as a unified object that reacts *directly* to the posited properties in the world. Any system that represents itself in such a way is bound to believe that there are properties of the world given directly to the object, which it takes to be itself. In subpersonal terms this object and these properties, as well as their relation to each other, are postulated entities that explain the sensory input. For instance, the fact that others talk about the system as someone with beliefs and desires (which is rooted in the same principle) can be explained by predicting itself in the same way.
2. **These properties are irreducible to physical, mechanical phenomena.**
   Since the generative model does not depict these properties as built up from simpler ones, but simply posits them to predict lower-level patterns, these properties don't seem (to the system) to be reducible to other properties.
3. **These properties are atomic, i.e., unstructured.**
   There are as many posited properties as there are distinct dispositions to be tracked. This also explains why one can learn to find structure in formerly unstructured qualia (cf. Dennett 1991, p. 49) once new discriminative behavior is learned.
4. **These properties are important to our lives/beings as humans/persons**
   This felt importance is obvious, given the putative role they play in the explanation provided by the generative model. These properties seem to be the causes of all our behavior: if one did not feel the painfulness of a pain, one would not scream; if one did not sense the funniness of a joke, one would not laugh, etc. Since the model is still needed for interacting with others, despite theoretical advances in the sciences this felt importance of qualia to our lives is very difficult to overcome.
5. **These properties are known to every living human being; it is not possible to sincerely deny their existence**
   This is due to the fact that our brains predict the behavior of others via a model that posits direct interaction between "agents" and first-order, non-relational object properties—the entities that are then named "qualia".

This list has considerable overlap with lists of features ascribed to qualia (e.g., Metzinger 2003, p. 68; Tye 2013), lending support to the thesis that we don't need a revolution in science to accommodate qualia, but rather a change in perspective: we might look at the creatures described above and see that "[t]hey are us" (Dennett 2000, p. 353).

5 Conclusion

I have given an interpretation of Dennett's theory of why there seems to be something more to consciousness than science can explain. My aim was to thereby address crucial questions, while sticking as closely to Dennett's philosophy as possible. The answer is a just-so story that shows how (plausible) selection pressures lead to beings that cannot help but believe that they are *more* than just "moist robots" (Dennett 2013a, p. 49)—because some important entities seem to be missing from the scientific description.

This story answers the questions why and how beings like us monitor their dispositions, and how this ability could have evolved. It also offers an answer as to why we don't recognize them as representations of our dispositions and why qualia are unlike other theoretical entities in that they are important for what we consider ourselves to be. The notion of an intermediate between first- and second-order intentional systems was introduced as a new conceptual instrument for satisfying the acquisition constraint and to lay the fundaments for the belief in mind-independent simple properties that dir-

ectly cause the behavior of agents. This in turn is the basis for the belief in qualia as intrinsic properties of experience.

This story might not provide an "insight into necessity" (cf. Dennett 1991, p. 401), but I am happy if it contributes to showing and clarifying a possibility: although it may *seem* that our best hypothesis for accounting for our belief in qualia is that they actually exist, this hypothesis might still be explained away.

Acknowledgements

I want to thank Thomas Metzinger and Jennifer Windt for the unique opportunity to participate in this project. I am also very grateful for the helpful remarks they and two anonymous reviewers gave to an earlier version of this paper.

References

Bechtel, W. (1985). Realism, instrumentalism, and the Intentional Stance. *Cognitive Science*, *9* (4), 473-497. 10.1207/s15516709cog0904_5

Blanke, O. & Metzinger, T. (2009). Full-body illusions and minimal phenomenal selfhood. *Trends in Cognitive Sciences*, *13* (1), 7-13. 10.1016/j.tics.2008.10.003

Botvinick, M. & Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature*, *391* (756). 10.1038/35784

Clark, A. (1996). Linguistic anchors in the sea of thoughts. *Pragmatics & Cognition*, *4* (1), 93-103. 10.1075/pc.4.1.09cla

——— (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36* (3), 181-204. 10.1017/S0140525X12000477

——— (2015). Embodied prediction. In T. Metzinger & J. W. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Dennett, D. C. (1971). Intentional systems. *Journal of Philosophy*, *68*, 87-106.

——— (1984). *Elbow room. The varieties of free will worth wanting.* Oxford, UK: Clarendon Press.

——— (1987). *The Intentional Stance.* Cambridge, MA: MIT Press.

——— (1991). *Consciousness explained.* New York, NY: Back Bay Books/Little, Brown and Company.

——— (1994). The Role of Language in Intelligence. In J. Khalfa (Ed.) *What is Intelligence? The Darwin College Lectures.* Cambridge, UK: Cambridge University Press. 10.1075/pc.4.1.09cla

——— (1995). *Darwin's dangerous idea: Evolution and the meanings of life.* New York, NY: Simon Schuster Paperbacks.

——— (1998a). Self-portrait. *Brainchildren: Essays on designing minds* (pp. 355-366). Cambridge, MA: MIT Press.

——— (1998b). Real patterns. *Brainchildren: Essays on designing minds* (pp. 95-120). Cambridge, MA: MIT Press.

——— (1999). The zombic hunch: Extinction of an intuition. *Royal Institute of Philosophy Millennial Lecture*

——— (2000). With a little help from my friends. In D. Ross, A. Brooks & D. Thompson (Eds.) *Dennett's Philosophy: A Comprehensive Assessment* (pp. 327-388). Cambridge, MA: MIT Press.

——— (2004). *Freedom Evolves.* London, UK: Penguin Books.

——— (2005). *Sweet dreams: Philosophical obstacles to a science of consciousness.* Cambridge, MA: MIT Press.

——— (2006). *Breaking the spell. Religion as a natural phenomenon.* New York, NY: Penguin.

——— (2009). Intentional Systems Theory. In B. P. McLaughlin, A. Beckermann & S. Walter (Eds.) *The Oxford handbook of philosophy of mind* (pp. 339-349). Oxford, UK: Oxford Handbooks Online.

——— (2010). The evolution of why. In B. Weiss & J. Wanderer (Eds.) *Reading Brandom: On making it explicit* (pp. 48-62). New York, NY: Routledge.

——— (2013a). *Intuition pumps and other tools for thinking.* New York, NY: W. W. Norton & Co..

——— (2013b). Expecting ourselves to expect: The Bayesian brain as a projector. *Behavioral and Brain Sciences*, *36* (3), 29-30. 10.1017/S0140525X12002208

——— (2015). Why and how does consciounsess seem the way it seems? In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Friston, K., Mattout, J. & Kilner, J. (2011). Action understanding and active inference. *Biological Cybernetics*, *104* (1-2), 137-160. 10.1007/s00422-011-0424-z

Friston, K. J., Shiner, T., FitzGerald, T., Galea, J. M., Adams, R., Brown, H., Dolan, R. J., Moran, R., Stephan, K. E. & Bestmann, S. (2012). Dopamine, affordance and active inference. *PLoS Computational Biology*, *8* (1), e1002327-e1002327. 10.1371/journal.pcbi.1002327

Gibson, J. J. (1986). *The ecological approach to visual perception.* Hillsdale, NJ: Erlbaum.

Harkness, D. (2015). From explanatory ambition to explanatory power-A commentary on Jakob Hohwy. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Hohwy, J. (2013). *The predictive mind.* Oxford, UK: Oxford Univeristy Press.

——— (2014). The self-evidencing brain. *Noûs*, *online.* 10.1111/nous.12062

——— (2015). The neural organ explains the mind. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Hohwy, J., Ropstorff, A. & Friston, K. (2008). Predictive coding explains binocular rivalry: An epistemological review. *Cognition*, *108* (3), 687-701. 10.1016/j.cognition.2008.05.010

Hume, D. (1995). *An inquiry concerning human understanding.* London, UK: Pearson.

Humphrey, N. (1983). *Consciousness regained.* Oxford, UK: Oxford University Press.

Lurz, R. W. (2010). Belief attribution in animals: On how to move forward conceptually and empirically. *Review of Philosophy and Psychology*, *2* (1), 19-59. 10.1007/s13164-010-0042-z

Madary, M. (2015). Extending the explanandum for predictive processing-A commentary on Andy Clark. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Marticorena, D. C.W., Ruiz, A. M., Mukerji, C., Goddu, A. & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, *14* (6), 1467-7687. 10.1111/j.1467-7687.2011.01085.x

McKay, R. T. & Dennett, D. C. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, *32*, 493-561. 10.1017 / S0140525X09990975

Metzinger, T. (2003). *Being no one. The self-model theory of subjectivity.* Cambridge, MA: MIT Press.

Roskies, A. (2015). Davidson on believers: Can non-linguistic creatures have propositional attitudes? In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Sellars, W. (1963). *Science, perception and reality.* London, UK: Routledge & Kegan Paul Ltd..

Seth, A. (2015). The cybernetic bayesian brain: From interoceptive inference to sensorimotor contingencies. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Stevens, G. C. (1989). The latitudinal gradients in geographical range: How so many species co-exist in the tropics. *American Naturalist*, *133* (2), 240-256.

Tye, M. (2013). Qualia. In E. N. Zalta (Ed.) *The Stanford Encyclopedia of Philosophy (Fall 2013 Edition)*

Wiese, W. (2015). Perceptual presence in the Kuhnian-Popperian Bayesian brain—A commentary on Anil Seth. In T. Metzinger & J. W. Windt (Eds.) *Open MIND.* Frankfurt a. M., GER: MIND Group.

Yu, P. & Fuller, G. (1986). A critique of Dennett. *Synthese*, *66* (3), 453-476. 10.1007/BF00414062

How our Belief in Qualia Evolved, and Why We Care so much

A Reply to David H. Baßler

Daniel C. Dennett

David Baßler's commentary identifies five unasked questions in my work, and provides excellent answers to them. His explanation of the gradual evolution of higher-order intentionality via a Bayesian account leads to an explanation of the persistence of our deluded belief in qualia.

Keywords
Belief in belief | Dispositions | Intentional systems | Qualia

Author

Daniel C. Dennett
daniel.dennett@tufts.edu
Tufts University
Medford, MA, U.S.A.

Commentator

David H. Baßler
davidhbassler@gmail.com
Johannes Gutenberg-Universität
Mainz, Germany

Editors

Thomas Metzinger
metzinger@uni-mainz.de
Johannes Gutenberg-Universität
Mainz, Germany

Jennifer M. Windt
jennifer.windt@monash.edu
Monash University
Melbourne, Australia

David Baßler's commentary is a model of constructive criticism, not only pointing to weaknesses but offering persuasive repairs. I have just two points of minor correction to offer before turning to my understanding of his interesting proposals for extensions to my view, which I am inclined to adopt.

First, then, the quibbles. I am happy to see him endorsing my frequent tactic of asking not how to explain x but rather asking how to explain why we believe in x in the first place, but I think that this is a procrustean bed on which to stretch my concept of intentional systems. In Dennett (1971) I was indeed offering an account of intentionality that was *demoting*, in that intentionality was not seen as a feature that sundered the universe into the mental and physical (as Brentano and others had claimed), but I don't like to think of it as dismissing intentionality as a real phenomenon—though of course many have interpreted me that way. Dennett (1991) tried to correct that misconstrual, showing that the phenomena of intentionality are real in their own way—any beings that don't discover these patterns are missing something important in the world. That aside, I

love the use he makes of Hume on miracles to introduce his treatment of our minds as witnesses, just not very good witnesses; their testimony can be explained in ways that do not grant the truth of some of their most cherished claims. As he puts it, the assumption of phenomenal consciousness "is deeply misleading because it makes us look for the wrong things, namely, the objects our judgments are about, rather than the causes of these judgments, which are nothing like these objects" (Baßler this collection, p. 2).

My other quibble is a similar elision I want to resist. He says: "Large parts of Breaking the Spell are dedicated to making understandable how 'belief in belief' could have evolved over the centuries, beginning long before the appearance of any religion" (Baßler this collection, p. 4). This misidentifies higher order belief, beliefs about beliefs, with belief in belief. The former did indeed evolve gradually over the eons, and I find Baßler's "just so story" about this gradual process enticing indeed, and will have more to say about it below, but belief in belief is a much younger (and almost always pernicious) phenomenon, which involves the deeply confused judgment that it is morally obligatory to try to get yourself to believe traditional nonsense when you know better. "If you don't believe in God, you are immoral. Therefore you must strive to believe in God. Belief in God is a good thing to inculcate in our children and in ourselves." Belief in belief didn't arrive on the human scene until the proto-religions (which originally had no need for the concept) hit upon this obligation as a way of protecting their hegemony against the lures of competing dogmas. Some proto-religions were blithely ecumenical, adopting the gods and demons of their neighbors' creeds as just another bit of lore about the big wide world, but this credulity could not long stand in the face of market competition and growing common knowledge about the objective world. Since many—probably most—people in the world now see through at least most of the nonsense, their persistent belief in belief is now a deplorable anachronism, a systematic source of hypocrisy. (A delightful cartoon in a recent New Yorker perfectly captures this folly. Two armies confront each other, flying identical banners; one mounted warrior says "There can be no peace until they renounce their Rabbit God and accept our Duck God.")

As I say, these are quibbles I have to get off my chest. Now to Baßler's substantive proposals. He organizes his commentary around five questions he says I haven't properly asked, and he has answers to all of them. He's right that these are gaps in my account. (1) Why do we need to monitor our dispositions? (2) How is self-monitoring accomplished? (3) How did this self-monitoring evolve in a *gradual* fashion? (4) Why do we misidentify our dispositions? (5) Why are we so attached to the idea of qualia?

His answers are constructed by taking on, for the sake of argument, my Intentional Systems Theory, and he gets it right, in all regards. Intentional Systems Theory (IST) presupposes, tactically, that any entity treated as an intentional system "is optimally designed to achieve certain goals. If there are divergences from the optimal path, one can, in a lot of cases, correct for this by introducing abstract entities or false beliefs." IST is, as I say, a competence model that leaves implementation or performance questions unaddressed.[1]

Then comes Baßler's major novelty: the idea of an intermediate competence between mere first-order intentional systems—which have no beliefs about beliefs (their own or others')—and full-fledged second-or-higher-order intentional systems—which can iterate the belief context. Such entities he calls (what else?) 1.5th order intentional systems (shades of David Marr's 1982 two-and-a-half-D sketch!). This is proposed to answer his first and second questions with a plausible and in principle testable evolutionary hypothesis. A system with only 1.5th order intentionality "is able to ascribe desires only in a very particular and concrete

1 In this regard it is strikingly similar to the free energy principle as presented by Hohwy (this collection); both use the assumption of biofunctional optimizing as an interpretive lever to make sense of the myriad complexities of the brain, assigning to the brain a fundamental task of acquiring accurate anticipations of the relevant causes in the organism's world. I have not yet been able to assess the costs and benefits of these two different ways of thinking of brains as future-producer: both are abstract, both court triviality if misused. This is a good topic for future work.

manner, i.e., actions that the object in question wants to perform with certain particular existing objects, that the system itself [the ascriber] knows about" (Baßler this collection, p. 6). He is wise to choose basic desires (for food, mating opportunities, safety, . . .) as the intentional states ascribed in this precursor mentality, since they are so readily "observable" in the immediate behavior of the object, giving our pioneer mind-reader a quick confirmation that it's on the right track, a small, gradual step for a Bayesian brain.

Now what selection pressures would favor such systems evolving gradually from mere first-order systems? To the primitive first-order systems, "the behavior of their conspecifics is unexplained noise to them." But then they make some simple discoveries. When they see an apple tree, they approach it, *and so do their conspecifics.* If they see a predator, they run, as do their kin. "One might indeed say that the desires of the agents are projected onto the world", Baßler says. Then, in a very substantive footnote that I wish were in the text—his footnotes contain much of value, and should not be passed over!—he adds: "What I mean by 'project' is that instead of positing an inner representation . . . . whose function is a desire, along with correct beliefs about the current situation, what is posited is an eat-provocative property of the apple itself. Both theoretical strategies allow for the prediction of the same behavior. The crucial difference is that attributing new properties to objects that are already part of the model is a simpler way of extending the model than positing a complex system of internal states to each agent" (Baßler this collection, p. 7, footnote 9). This answers question (3).

He then imagines, plausibly, that these 1.5^{th}-order systems will evolve a system of communication, but this (as I and others have argued) necessarily involves hiding information from others, which involves having an internal cache of self-monitored knowledge one can choose to divulge or not, depending on circumstances. And this in turn—Baßler's next major innovation—leads them to become "Agents [who] believe in the existence of a special kind of properties: they believe that they approach apples because they are sweet, cuddle babies because they are cute, laugh about jokes because they are funny." This primitive concept of causation serves them well, of course, and is just the sort of simplification to expect in a Bayesian brain, answering question (4).

Now for the icing on the cake, Baßler's answer to question (5) about why we care about qualia. As he notes, "It is not obvious why we do not react as disinterestedly to their denial as we did to the revelation that there is no ether" (Baßler this collection, p. 5). Here is his explanation: science comes along and starts to dismantle the handy manifest image, with all its Gibsonian affordances, and for those creatures capable of understanding science, a new problem arises: something is being taken away from them! All those delectable properties (and the abhorrent properties as well, of course). Philosophers "still see that there is something missing, and since cuteness is not a property of the outside world, they conclude that it must be a property of the agents themselves" (Baßler this collection, p. 8). "We have the zombic hunch because it seems to us that there is something missing and it seems so because our generative models are built on the assumption that there are properties of things out there in the world to which systems like us react in certain ways. . . . We dismiss robots because we know they can only react to measurable properties, which do not seem to us to be the direct cause of our behavior" (ibid.).

This rings true to me, and I hadn't seen this way of accounting for the persistence of the zombic hunch. Baßler proposes that "the reason we, intuitively, do not accept a robot as a subject like ourselves is because we know how the robot does it; we know that it calculates, maybe even in a PP manner–we know that it does not react directly to the properties that seem to exist and that seem to count" (this collection, p. 9). He goes on to list five further features his account provides for. The properties we delusionally persist in "projecting" as qualia are (1) "'given directly' to a person", (2) "irreducible to physical, mechanical phenomena", (3) "atomic, unstructured", (4) "important to our lives/beings as humans/persons", and (5)

"known to every living human being; it is not possible to sincerely deny their existence" (Baßler this collection, p. 9). I particularly like the way that his account explains why (4) is a feature: "These properties seem to be the causes of all our behavior: if one did not feel the painfulness of a pain, one would not scream; if one did not sense the funniness of a joke, one would not laugh, etc. Since the model is still needed for interacting with others, despite theoretical advances in the sciences this felt importance of qualia to our lives is very difficult to overcome" (Baßler this collection, p. 9).

I see that my response consists in large measure of approving quotations from Baßler's commentary! But that is as it must be; I want to confirm in detail and acknowledge the nice way his proposals dovetail with my account, expanding it into new territory, and helping me see what I have so far only dimly appreciated: just how valuable the new Bayesian insights are.

But let me end with a friendly amendment of my own. Baßler's interpretation of my view is at one point a simplification, probably just for gracefulness of exposition, and perhaps meant itself as a friendly amendment, but I want to issue a caveat. Baßler takes me to be saying that, for such properties as cuteness and color, "we misidentify dispositions of the organism with properties of another object" (this collection, p. 3) and goes on to have me holding that "This means, under a personal level description, that we believe that there are properties independent of the observer, such as the cuteness of babies, the sweetness of apples, or the blueness of the sky" (ibid., p. 4). I want to put this slightly differently. It is not that there is nothing objective about babies that makes them cute (or of the sky that makes it blue) but just that these objective, observer-independent properties are themselves curiously dispositional: they are, as he notes at one point, what I have called "lovely" properties. They can only be *defined* relative to a target species of observers, such as normally sighted—not "color-blind"—human beings, as contrasted with tetrachromats such as pigeons, for instance. But their existence as properties is trivially objective and observer-independent. Thus rubies were *red* before color vision evolved on this planet in the sense that if a time machine could take normal human beings back to the early earth, they would find rubies to be red. And some strata exposed by primordial earthquake faults would have been *visible,* to some kinds of eyes and not to others. Probably dinosaur babies were cute, since, as John Horner (1998) has argued, evidence strongly suggests that they were altricial, requiring considerable parental attention, and having the foreshortened skull and facial structure of prototypically cute juvenile animals, including birds. The science-endorsed properties, both external and internal, are so hugely different from what the manifest image makes them out to be, that it is a pickwickian stretch to say that science has discovered "what cuteness is" or "what color is," but it is also deeply misleading to say that science has discovered that nothing is cute, or colored, after all. And so in a similar vein, I have to contend with how to occupy the awkward middle ground between denying that there are qualia at all, or saying that qualia are something real, but something utterly unlike what most people *think* (and philosophers *say*) qualia are.

1 Conclusion

Baßler has provided me with a plausible and testable extension of my Intentional System Theory with his innovation of a 1.5^{th}-order intentional system, showing in outline how higher-order intentional systems might evolve from their more primitive ancestors. And he has also shown new ways of explaining a point that many people just cannot get their heads around. As my former student Ivan Fox (1989) once put it, "Thrown into a causal gap, a quale will simply fall through it." See also Fox's essay, "Our Knowledge of the Internal World" (1994) and my commentary on it (1994), which I discovered, on rereading just now, to be groping towards some of the points in Baßler's commentary. I challenged Ivan Fox to "push further into the engineering and not just revel in the specs" (Dennett 1994, p. 510), and Baßler has done just that.

References

Baßler, D. H. (2015). Qualia explained away: A commentary on Daniel C. Dennett. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Dennett, D. C. (1971). Intentional systems. *The Journal of Philosophy*, *68* (4), 87-106. 10.2307/2025382

——— (1991). Real patterns. *The Journal of Philosophy*, *88* (1), 27-51. 10.2307/2027085

——— (1994). Get real. *Philosophical Topics*, *22* (1 & 2), 59-106.

Fox, I. (1994). Our knowledge of the internal world. *Philosophical Topics*, *22* (1 & 2), 59-106.

Hohwy, J. (2015). The neural organ explains the mind. In T. Metzinger & J. M. Windt (Eds.) *Open MIND*. Frankfurt a. M., GER: MIND Group.

Horner, J. R., Gorman, J., Henderson, D. & Blumer, T. L. (1998). *Maia: A dinosaur grows up.* Bozeman, MT: Museum of the Rockies, Montana State University.

Marr, D. (1982). *Vision.* New York, NY: Freeman.